2017 American Control Conference
Sheraton Seattle Hotel
May 24–26, 2017, Seattle, USA

Anomalous Node Detection in Networks with Communities of Different Size

Juan Campos and Jorge Finke[1]

Abstract— **Based on two simple mechanisms for establishing and removing links, this paper defines an event-driven model for the anomalous node detection problem. This includes a representation for (i) the tendency of regular nodes to connect with similar others (i.e., establish homophilic relationships); and (ii) the tendency of anomalous nodes to connect to random targets (i.e., establish random connections across the network). Our approach is motivated by the desire to design scalable strategies for detecting signatures of anomalous behavior, using a formal representation to take into account the evolution of network properties. In particular, we assume that regular nodes are distributed across two communities (of different size), and propose an algorithm that identifies anomalous nodes based on both geometric and spectral measures. Our focus is on defining the anomalous detection problem in a mathematical framework and to highlight key challenges when certain topological properties dominate the problem (i.e., in terms of the strength of communities and their size).**

I. INTRODUCTION

The lofty aim of network models is to serve as analytical frameworks that capture the dynamic relationships across large interconnected systems. It is of interest to understand how interaction processes explain the formation of structure, i.e., how mechanisms for establishing and removing links influence the evolution of topological properties. Mechanism-based models provide the basis for the design of algorithms that take account of regular patterns in networks.

A common approach to the anomalous node detection problem is to study the evolution of local and global properties, including (i) the proportion of close-knit groups (i.e., subgraphs of k nodes, each with at least $k/2$ neighboring nodes) [1], [2]; and (ii) the formation of communities (i.e., groups of nodes with tight connections within and sparse connections across them) [3], [4]. How to detect close-knit groups of anomalous nodes on networks with different-sized community structures remains an open challenge.

The contribution of this paper is twofold. First, we introduce a model based on two mechanisms, which characterizes how regular nodes impact the size and strength of communities. Second, we propose an anomalous node detection algorithm that combines geometric and spectral network measures. As in [5], our approach aims to effectively attribute detection signatures to patterns resulting from nodes that persistently engage in random link attacks (RLAs) [6]. Unlike the work in [5], the design of our algorithm is based on a representation of interactions underlying the behavior of regular nodes. We take a discrete-event modelling approach and use simulations to give insight into scenarios where the challenge of how to detect anomalous nodes is significant. Our results suggest that the ability to detect anomalous nodes is highly constrained by the degree to which homophilic relationships impact community strength. The formation of strong communities of similar size facilitates the detection of anomalous nodes.

II. PRELIMINARIES

A. Notation

Let $G = (G(0), G(1), ...)$ represent a sequence of unweighted, undirected networks. Each network $G(t) = (N, A(t))$ is composed of a set of nodes $N = \{1, ..., n\}$ and a set of edges $A(t)$. An element $\{i, j\} \in A(t)$ if and only if node i links to node j, and $\{i, i\} \notin A(t)$ for all $i \in N$. Note that the set of nodes N remains constant. It is composed of anomalous nodes (referred to as nodes of type 0) and two types of regular nodes (referred to as nodes of type 1 and 2). The function $g : N \to \{0, 1, 2\}$ defines the type of a node. Let $N_\delta = \{i \in N : g(i) = \delta\}$ be the set of nodes of type δ, and $n_\delta = |N_\delta|$ the size of N_δ. Assume that $n_2 \geq n_1$, so that N_2 refers to the majority group whenever there exists a difference in group size. Let $A_i(t) = \{\{j', j\} \in A(t) : j' = i\}$ be the set of edges that link node i to its neighboring nodes, and $A_i^c(t)$ denote the complement of $A_i(t)$. Furthermore, let $k_i(t) = |A_i(t)|$ denote the number of neighbors of node i, and $k_i^\delta(t) = |\{\{i, j\} \in A_i(t) : g(i) = g(j)\}|$ the number of same-type neighbors of node i. Moreover, at any time t let $R_i(t) \subseteq A_i(t)$ be a subset of edges that node i is able to redirect. Consider the following assumption.

A1 Suppose that $|R_i(t)| = |R_i(0)| = |R_j(0)| = r$ for some constant $r \in \mathbb{N}$, and $R_i \cap R_j = \emptyset$ for all $t \geq 0$ and $i, j \in N$.

Assumption A1 requires that all nodes redirect the same number of edges. Moreover, each edge is redirected by a unique node at any time. Based on assumption A1, Section III describes decision-making mechanisms that encourage regular nodes to connect with other nodes of the same type, contributing to the formation of communities. In contrast, the behavior of anomalous nodes is characterized by weak degrees of membership to any particular community, resulting from the following generic behavior.

Definition 1: Random links attacks (RLAs) are a collaborative action by a close-knit group of anomalous nodes, which target randomly selected regular nodes, with no particular preference for any type of node [6]. An anomalous

[1] Both authors are with the Department of Electrical Engineering and Computer Science, Pontificia Universidad Javeriana, Cali, Colombia. juan.campos, finke@ieee.org

node establishes a link with a regular node with probability w_a (and a link with another anomalous node with probability $1 - w_a$).

In online social networks, such as Facebook, Google+ or Twitter, an example of these types of attacks include groups of fake interconnected users who randomly connect to other users.

B. Geometric measures

Next consider the following two definitions. The cohesion index is a distance function, defined based on the number of same-type nodes in a neighborhood of radius 1.

Definition 2: The cohesion index for the group of nodes of type δ is $h_\delta(t) = \frac{1}{n_\delta} \sum_{i \in N_\delta} \frac{k_i^\delta(t)}{k_i(t)}$, where $\frac{k_i^\delta}{k_i}$ represents the proportion of neighboring nodes that are of the same type.

Definition 3: The modularity of the network $G(t)$ is

$$q(t) = \sum_{\delta=1}^{2} \Bigg(\frac{|\{i,j\} \in A(t) : g(i) = g(j) = \delta|}{|A(t)|} - \frac{|\{i,j\} \in A(t) : g(i) = \delta \vee g(j) = \delta|^2}{|A(t)|^2} \Bigg)$$

Here, the modularity of the network measures the strength of community formation based on the number of edges between nodes of the same type compared to the number of edges between nodes of different type. It captures the idea of a community as a group of nodes with higher inter- than intraconnectivity [3]. Modularity values above 0.3 indicate that the network exhibits a well-defined community structure [7]. In particular, note that both the modularity, and the cohesion indices depend on the number of links established between nodes of the same type. Unlike the cohesion index, the modularity yields a single value for the entire network.

C. Spectral measures

An alternative way to describe topological properties is by using spectral measures. To define them, it is convenient to refer to $G(t)$ through its adjacency matrix $M(t)$, where $m_{ij}(t) = m_{ji}(t) = 1$ if $\{i, j\} \in A(t)$, and $m_{ij}(t) = 0$ otherwise. Since $M(t)$ is symmetric, all its eigenvalues are real numbers. Let $\lambda_1(t) \geq \lambda_2(t) \geq \ldots \geq \lambda_n(t)$ be the eigenvalues associated to $M(t)$ at time t, and let $z_j(t)$ represent the eigenvector associated to the eigenvalue λ_j, where z_{ji} is the i-th element of z_j. Moreover, $z_i'(t) = (z_{1i}(t), z_{2i}(t), \ldots, z_{ni}(t))$ represent the spectral coordinates associated to node i. When a network has two well-defined communities, a common approach to approximate the spectral coordinates of node i is to focus the first two entries of $z_i'(t)$. Consider the following measure associated to a node.

Definition 4: The node-non-randomness of node i represents the sum of non-randomness values of all its edges, and it is denoted by

$$f_i(t) = \sum_{j=1}^{2} \lambda_j(t) \; z_{ji}^2(t) \qquad (1)$$

Based on eq (1), consider the following algorithm (introduced in [5]) to identify a set of potential anomalous nodes, denoted by V_s.

Algorithm 1 Identifying potential anomalous nodes (suspects) based on node-non-randomness

```
Input: Adjacency matrix $M(t)$
Output: Set of suspects $V_s(t)$
Calculate $\lambda_j(t)$ and $z_j(t)$
for $i \leftarrow 1$ to $n$ do
    Calculate the node-non-randomness $f_i(t)$.
    Calculate $B_i^E(t)$ and $B_i^V(t)$.
    if $f_i \leq B_i^E(t) + \alpha(B_i^V(t))^{1/2}$ then
        $V_s(t) \leftarrow V_s(t) \cup \{i\}$
    end if
end for
return $V_s(t)$
```

A node is considered a potential anomalous node (called a suspect) if

$$f_i \;\leq\; B_i^E + \alpha(B_i^V)^{1/2} \qquad (2)$$

where $\alpha = 2$ is a design parameter, and

$$B_i^E = k_i^2 \sum_{j=1}^{2} \frac{E[x_j]^2}{\lambda_j} + \frac{k_i}{n}\Big(1 - \frac{k_i}{n}\Big) \sum_{j=1}^{2} \frac{1}{\lambda_j}$$

and

$$B_i^V = \frac{4k_i^3}{n}\Big(1 - \frac{k_i}{n}\Big) \sum_{j=1}^{2} \frac{E[x_j]^2}{\lambda_j^2} + \frac{2k_i^2}{n^2}\Big(1 - \frac{k_i}{n}\Big)^2 \sum_{j=1}^{2} \frac{1}{\lambda_j^2}$$

are the upper bounds of the expected value and variance of f_i, respectively.

D. Performance measures

Nodes identified as anomalous nodes are a subset of the set of suspect nodes V_s, denoted by V_a. If $i \in V_a \subseteq V_s \subseteq N$, then node i is considered an anomalous node. To evaluate the performance of detecting anomalous nodes, consider the following two measures.

Definition 5: The false positive error rate is the number of regular nodes reported as being anomalous, over the total number of reported nodes, i.e., $e_1 = |\{i \in N_1 \cup N_2 : i \in V_a\}|/|V_a|$.

Definition 6: The true positive error rate is the number of anomalous nodes that are reported over the total number of anomalous nodes, i.e., $e_2 = |\{i \in N : i \in V_a \wedge i \in N_0\}|/n_0$.

The performance of a detection algorithm is considered acceptable if $e_1 \leq 0.05$ and $e_2 \geq 0.95$.

III. THE NETWORK MODEL

Two simple mechanisms for establishing and removing links drive the evolution of the network. In particular, every time index t, a randomly selected node i redirects links based on the following mechanisms.

M1 Link connections: If node i is a regular node, i.e., $g(i) \in \{1, 2\}$, it establishes a new edge to node c, $c \neq i$, such that $\{i, c\} \notin A_i$. The probability that node i links to node c at time t depends on both the type and the degree of node c, and is given by

$$\pi_c(t) = w_c\, k_c(t)\, \frac{1}{\sum\limits_{\{i,j\} \in A_i^c(t)} w_j k_j(t)}$$

with

$$w_c = \begin{cases} w & \text{if } g(i) = g(c), \\ 1 - w & \text{if } g(i) \neq g(c). \end{cases}$$

where k_c represents the degree of node c, and w the preference of regular nodes to associate with similar others.

If node i is an anomalous node, i.e., $g(i) = 0$, then it engages in a RLA (based on Definition 1). The probability that node i establishes a new link with node c is given by

$$\pi_c = \begin{cases} \dfrac{1 - w_a}{n_0 - k_i^\delta(t)} & \text{if } g(c) = 0, \\[2ex] \dfrac{w_a}{n_1 + n_2 - (k_i - k_i^\delta(t))} & \text{if } g(c) \in \{1, 2\}. \end{cases}$$

M2 Link disconnections: Node i removes the link to some node $d \neq c$, $\{i, d\} \in R_i(t)$, selected according to a uniformly random distribution.

Note that mechanism M1 promotes connections between regular nodes of the same type (it captures the proposition that individuals with shared interest or hobbies tend to have a higher density of links). Moreover, nodes are more likely to establish links with nodes with a higher degree. Based on mechanism M1, it is natural to expect a greater number of attacks against the majority group. Mechanism M2 forces node i to remove a link from one of its neighbors, regardless of the type of node i. Together mechanisms M1 and M2 guarantee that the number of edges that a node is able to redirect remains constant.

Let $k^\delta(t) = [k_1^\delta(t), ..., k_n^\delta(t)]^\top$ be the number of same-type neighbors and $k(t) = [k_1(t), ..., k_n(t)]^\top$ be the total number of neighbors for each node. Furthermore, let $x(t) = \left[k^\delta(t)^\top, k(t)^\top\right]^\top$ be the state of the network at time t.

An event e_i occurs at time t, if node i redirects one of its links according to mechanisms M1 and M2. More formally, an event e_i occurs if $e_i \in g_e(x)$, where $g_e(x)$ denotes a function that enables an event. If $e_i \in g_e(x)$, then the next state of the network $x(t+1)$ is defined by $x(t+1) = f_e(x(t))$, where f_e is an operator defined by the following state transitions.

If $g(i) = g(d) = g(c)$:

$$k_i^\delta(t+1) = k_i^\delta(t)$$
$$k_c^\delta(t+1) = k_c^\delta(t) + 1$$
$$k_d^\delta(t+1) = k_d^\delta(t) - 1$$

If $g(i) = g(d) \neq g(c)$:

$$k_i^\delta(t+1) = k_i^\delta(t) - 1$$
$$k_c^\delta(t+1) = k_c^\delta(t)$$
$$k_d^\delta(t+1) = k_d^\delta(t) - 1$$

If $g(i) \neq g(d) = g(c)$:

$$k_i^\delta(t+1) = k_i^\delta(t) + 1$$
$$k_c^\delta(t+1) = k_c^\delta(t) + 1$$
$$k_d^\delta(t+1) = k_d^\delta(t)$$

If $g(i) \neq g(c) = g(d)$ or $g(i) \neq g(c) \neq g(d) \neq g(i)$:

$$k_i^\delta(t+1) = k_i^\delta(t)$$
$$k_c^\delta(t+1) = k_c^\delta(t)$$
$$k_d^\delta(t+1) = k_d^\delta(t)$$

Moreover, it is always the case that

$$k_i(t+1) = k_i(t)$$
$$k_c(t+1) = k_c(t) + 1$$
$$k_d(t+1) = k_d^\delta(t) - 1$$

Note that the cohesion indices of any group at time t can be specified based on the state of the network (according to Definition 2).

Finally, we require that the network $G(0)$ satisfies the following assumptions.

A2 *Initial configuration:* Each set of regular nodes can show total cohesion ($r < n_1$); moreover, every node can connect to both types of nodes at the same time ($r \geq 2$).

A3 *Persistency of events:* If the cohesion index of the group N_δ satisfies $0 < h_\delta < 1$, then an event $e_i \in g(x)$ occurs.

For each group of regular nodes to be able to reach a cohesion index $h_\delta = 1$, assumption A2 restricts the maximum number of links that a node can establish. In particular, each node $i \in N_\delta$ can establish at most $\min\{n_\delta\} - 1$ links.

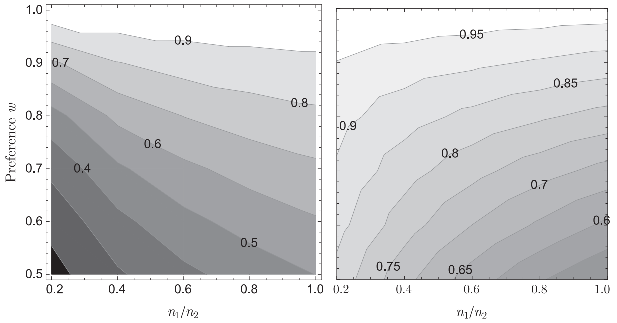
The following section describes geometric and spectral properties of the model.

IV. MODEL PROPERTIES

A. Geometric properties measures

To illustrate the effect of relative group size n_1/n_2 on the homophilic relationships of regular nodes, consider mechanisms M1 and M2 operating on an initial random network. Let $n = 10^4$ and $r = 30$. After $t = 10^5$, the cohesion indices for the minority and majority group reach stationary values, which are shown in Fig. 1. Note that the cohesion indices depend on the value of w. Varying the relative size between groups has a strong effect on the cohesion of both groups. Note also that for a fixed preference w, different groups can reach different levels of cohesion, depending on their relative size: For the majority group, the cohesion index decreases as the proportion n_1/n_2 increases; for the minority group the opposite is true, the cohesion index increases as n_1/n_2 increases.

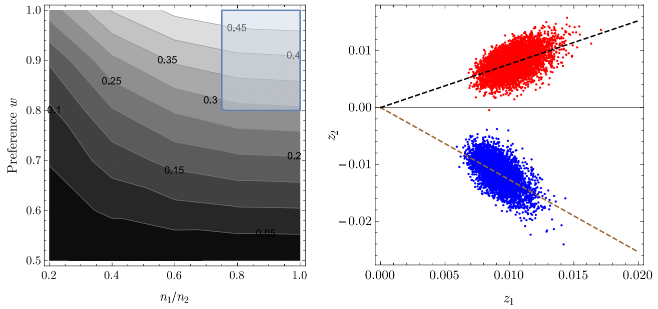
Next, Fig. 2(a) shows modularity resulting from varying relative sizes and preference levels. As expected, the modularity increases as preference levels increase. According to Definition 3, the model generates networks with well-defined community structures, i.e., $q \geq 0.3$, for $w \geq 0.8$ and $n_1/n_2 \geq 0.75$ (highlighted by the box in Fig. 2).



(a) Minority group.

(b) Majority group.

Fig. 1. Expected cohesion indices for the minority and majority groups.



(a) Modularity measure.

(b) Spectral coordinates of a generated network.

Fig. 2. Modularity and spectral coordinates.

B. Spectral measures

Fig. 2(b) shows the spectral coordinates for a network with $n_1/n_2 = 0.75$ and $w = 0.8$. As the preference to associate with similar others increases, two clusters of nodes are formed, and the projection of node coordinates are grouped along two quasi-orthogonal lines, which represent the division of the network into two communities.

V. Anomalous Node Detection

In [5] the authors show that the coordinates of the nodes engaging in RLAs asymptotically follow a multivariate normal distribution. We use a similar idea to detect potential anomalous nodes (suspects) described below as Algorithm 2. Under the assumption that anomalous nodes form close-knit groups (in an effort to masquerade as regular nodes), we also search the set of suspects for subgraphs with a high link density, described below as Algorithm 3.

A. Identifying suspects across the spectral space

The main idea behind Algorithm 2 is to exploit the fact that the spectral coordinates of anomalous nodes follow a normal distribution, which means that the expected value and variance satisfy the inequalities

$$E[z_{ji}(t)] \leq \frac{k_i(t)\, E[z_j(t)]}{\lambda_j(t)},$$

$$V[z_{ji}(t)] \leq \frac{k_i(t)}{n}\left(1 - \frac{k_i(t)}{n}\right)\frac{1}{\lambda_j(t)^2}.$$

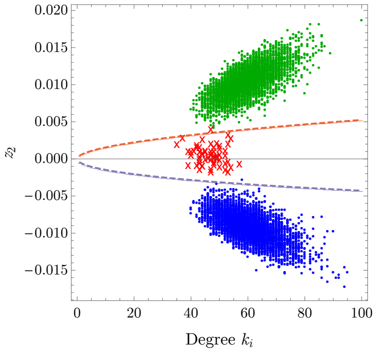


Fig. 3. Spectral coordinate: z_2 versus node degree. The dashed curves represent the upper and lower bounds on the expected values of the second coordinate of a anomalous node with degree k_i, $E[z_{2i}(t)]$.

It can be shown that when there is no collaboration between anomalous nodes (i.e., with $w_a = 1$), both expressions are satisfied with equality [5]. The upper bounds on the expected value and variance are

$$U^E_{ji}(t) = \frac{k_i(t)\, E[z_j(t)]}{\lambda_j(t)}, \tag{3}$$

$$U^V_{ji}(t) = \frac{k_i(t)}{n}\left(1 - \frac{k_i(t)}{n}\right)\frac{1}{\lambda_j(t)^2}. \tag{4}$$

Based on Algorithm 2, node i is reported as a suspect if all its spectral coordinates are within the confidence interval $E[z_{ji}] \pm \epsilon V(z_{ji})^{1/2}$, where $\epsilon > 0$ is a design parameter that denotes the $\frac{1+p}{2}$ quantile of the standard normal distribution. The larger the value of ϵ, the less likely that the algorithm fails to identify actual anomalous nodes as suspects (which leads to an increase the true positive error rate). However, the larger the value of ϵ, the more likely that Algorithm 2 identifies regular nodes as suspects (i.e., thereby increasing e_1). Here we choose $\epsilon = 2$, so that the confidence interval covers more than the 95%.

Fig. 3 shows the spectral coordinate of z_2 for regular and anomalous nodes as a function of node degree. Note that the component of the spectral coordinates of anomalous nodes (marked with crosses) are mostly located in the region between the bounds; regular nodes fall outside this envelope.

B. Detecting anomalous nodes

Based on the assumption that few regular nodes in the suspect set V_s have an high number of edges between them [5], Algorithm 3 identifies suspect nodes that are involved in dense subgraphs.

VI. Algorithm Performance

A. Static detection

Previous sections define three different algorithms that help to identify suspect and anomalous nodes by combining geometric and spectral measures. In particular, Algorithms 1 and 2 identify suspects, and Algorithm 3, applied on the resulting set of suspects, distinguishes anomalous from suspect

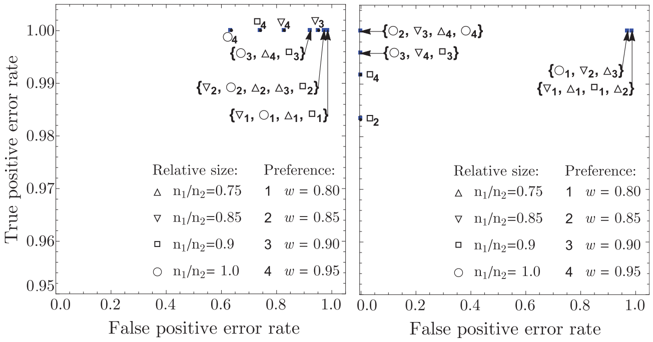
Algorithm 2 Identifying suspects nodes at time t

```
Input: Adjacency matrix $M(t)$
Output: Set of suspects $V_s(t)$
Calculate $\lambda_j(t)$ and $z_j(t)$
for $j \leftarrow 1$ to $2$ do
    for $i \leftarrow 1$ to $n$ do
        Calculate $U_{ji}^E(t)$ and $U_{ji}^V(t)$ using eq. (3) and (4).
        if $z_{ji}(t) \notin (U_{ji}^E(t) - \epsilon\ U_{ji}^V(t), U_{ji}^E(t) + \epsilon\ U_{ji}^V(t))$ then
            $V_s^j(t) \leftarrow V_s^j(t) \cup \{i\}$
        end if
    end for
end for
$V_s(t) \leftarrow V_s^1(t) \cap V_s^2(t)$
return $V_s(t)$
```

Algorithm 3 Identify anomalous nodes at time t

```
Input: Set of suspects $V_s(t)$
Output: Set of anomalous nodes $V_f(t)$
$G_s(t) \leftarrow$ subgraph formed by $V_s(t)$
$j \leftarrow$ node with the lowest degree in $G_s(t)$
$N_n \leftarrow$ number of nodes in $G_s(t)$
$N_e \leftarrow$ number of edges in $G_s(t)$
$\Delta^* \leftarrow N_e/N_n$
while $N_n > 1$ do
    $G_s(t) \leftarrow G_s(t)$ without $j$
    Calculate $N_e$ and $j$
    $N_n \leftarrow N_n - 1$
    $\Delta \leftarrow N_e/N_n$
    if $\Delta > \Delta^*$ then
        $\Delta^* \leftarrow \Delta$
        $V_f(t) \leftarrow$ Nodes of $G_s(t)$
    end if
end while
return $V_f(t)$
```

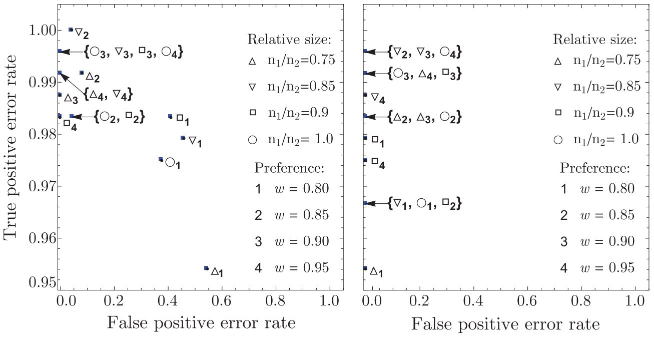
nodes. In this section, we compare the performance of Algorithm 1 followed by Algorithm 3 (introduced in [5], referred to as Approach A) with the performance of Algorithm 2 followed by Algorithm 3 (the proposed approach, referred as Approach B).



(a) Detection of suspects based on Algorithm 1.

(b) Detection of anomalous nodes based on Approach A.

Fig. 4. Detection based on Approach A.



(a) Detection of suspects based on Algorithm 2.

(b) Detection of anomalous nodes based on Approach B.

Fig. 5. Detection Approach B.

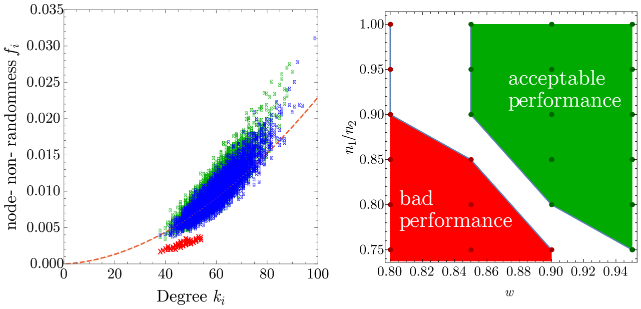
To measure the performance of both approaches we generate different networks with preference levels w varying from 0.8 to 0.95 and relative group sizes n_1/n_2 of 0.75, 0.85, 0.9, and 1.0. Note that the model requires a preference of at least $w = 0.8$ to generate a modularity greater than 0.3 (i.e., to generate a well-defined community structure). Moreover, consider $w_a = 0.7$, $n = 10^4$, and $n_0 = 60$. We apply the detection approaches at time $t = 10^5$, after the cohesion indices for both groups reach stationary values.

Figs. 4(a) and 4(b) show the performance of Algorithm 1 (by itself) and under Approach A. Each data point represents the average error rate of false positives e_1 and true positives e_2 for 4 simulation runs. For example, ∇_3 represents the mean performance for networks generated with $n_1/n_2 = 0.85$ and $w = 0.90$.

According to Fig. 4(a), note that Algorithm 1 yields a true positive error rate close to 1.0 for most network variations (i.e., it detects almost all anomalous nodes), but its false positive error rate exceeds 0.8 for half the scenarios. This means that a large number of regular nodes tend to be reported as potential anomalous nodes. In fact, Algorithm 1 identifies more than 20% of the nodes of the network as anomalous nodes, especially when the two communities are significantly different in size and the homophilic relationships are relatively weak (with a preference w lower than 0.90).

Figs. 5(a) and 5(b) illustrate the performance of Algorithm 2 (by itself) and under Approach B for the same networks. Note that Algorithm 2 makes an effective selection of suspects ($e_1 \leq 0.05$ and $e_2 \geq 0.95$) for almost all network variations, except for scenarios with relatively low preference levels ($w = 0.8$, which results in cohesion indices of 0.8 and 0.7 for the majority and the minority group, respectively). For networks with weak homophilic relationships, a similar performance is achieved when Algorithm 3 is applied to the set of suspects. Approach B shows that for all cases the performance yields a true positive error rate above 0.95, and a false positive error rate close to zero.

Note that Algorithm 3 is crucial to improve the performance of Approach A, which is not the case for Approach B. Fig. 6(a) suggests that the low performance of Approach A is due to the design parameter α (defined in eq. 2). Note that the detection bound is appropriate because the variable of node-non-randomness is a linear combination of non-central χ^2 random variables. However, the value of α is too sensitive to both group size and preference level.



(a) Node-non-randomness vs degree.

(b) Area of acceptable performance for Approach A.

Fig. 6. Suspect selection criterium and performance of Approach A.

Fig. 4(b) shows that Approach A yields an acceptable performance for some networks, while Approach B for all generated network variations. Fig. 6(b) illustrates the area of acceptable performance for Approach A as a function of the level of preference w and the relative group size n_1/n_2. The area of acceptable performance for Approach B covers all model parameters considered in Fig. 6(b).

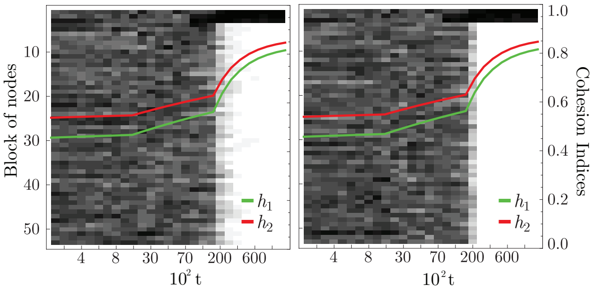
Finally, consider the boundary between the areas of acceptable and bad performance in Fig. 6(b) (denoted by the area in white). Note that high values of n_1/n_2 facilitate the detection of anomalous nodes for Approach A.

B. Dynamic detection

In the previous section we applied the detection algorithms after the cohesion indices reach stationary values. Next, we want to determine the time after which applying Algorithm 2 (by itself) or under Approach B would yield an acceptable performance. For this scenario, we analyze a network generated by the model with preference level $w = 0.85$ and relative group size $n_1/n_2 = 0.85$, for over 10^4 time steps. Moreover, as in the previous section we let $w_a = 0.7$, $n = 10^4$, and $n_0 = 60$.

Figs. 7(a) and 7(b) illustrate the dynamic behavior of 1000 regular nodes together with a set of anomalous nodes. The shade of each block represents the frequency with which a set of 20 nodes is reported as anomalous. Black blocks represent groups of nodes where each of them is reported as anomalous node, and the white blocks means that none of the 20 nodes is accused as an anomalous node. The first three blocks represent the actual set of nodes performing RLAs.

According to Figs. 7(a) and 7(b) the time indices from which an acceptable performance can be expected are $t = 500 \times 10^2$ for Algorithm 2 (by itself) and $t = 250 \times 10^2$ under Approach B. Fig. 7(b) shows that Approach B has an acceptable performance when the cohesion indices of both groups of regular nodes reach 80% of their final stationary value. Note that Algorithm 2 alone requires that these levels reach at least 90% of their final values (see Fig. 7(a)).



(a)Detection of suspects by Algorithm 2.

(b) Detection of anomalous nodes by Approach B.

Fig. 7. Dynamic detection.

VII. CONCLUSIONS AND FUTURE WORK

The proposed detection algorithm exploits the distribution of the spectral coordinates of the nodes, identifying anomalous nodes with a negligible false positive error rate for all network variations. Simulations show that the performance of previous approaches strongly depends on the relative difference in community size, and the level of preference of regular nodes to associate to similar others.

The network model also provides good insight into the challenges of dynamic detection of anomalous nodes. In particular, it serves as an analytical framework to establish detection thresholds for acceptable performance. Analyzing the behavior of the proposed algorithm over time, it suggests that cohesion indices of groups of regular nodes is a key criterion to determine the effectiveness of dynamic detection algorithms. Analyzing the behavior of the proposed approach when more than two groups make up the community structure of a network remains a future research direction.

ACKNOWLEDGMENTS

This research was carried out by the Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA). It is leading by the Pontificia Universidad Javeriana Colombia and it was funded by the Ministry of Information Technologies and Telecommunications of the Republic of Colombia (MinTIC) through the Colombian Administrative Department of Science, Technology and Innovation (COLCIENCIAS) within contract No. FP44842-anex46-2015.

REFERENCES

[1] P. Moriano and J. Finke, “Model-based fraud detection in growing networks,” in *Proceedings of the Conference on Decision and Control.* IEEE, 2014, pp. 6068–6073.

[2] C. C. Bilgin and B. Yener, “Dynamic network evolution: Models, clustering, anomaly detection,” *IEEE Networks*, 2006.

[3] A. Clauset, M. Newman, and C. Moore, “Finding community structure in very large networks,” *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 70, no. 6, p. 066111, 2004.

[4] R. Kaur and S. Singh, “A survey of data mining and social network analysis based anomaly detection techniques,” *Egyptian Informatics Journal*, vol. 17, no. 2, pp. 199 – 216, 2016.

[5] X. Ying, X. Wu, and D. Barbará, “Spectrum based fraud detection in social networks,” in *International Conference on Data Engineering*, 2011, pp. 912–923.

[6] N. Shrivastava, A. Majumder, and R. Rastogi, “Mining (social) network graphs to detect random link attacks,” in *2008 IEEE 24th International Conference on Data Engineering*, 2008, pp. 486–495.

[7] G. Csányi and B. Szendrői, “Structure of a large social network,” *Physical Review E*, vol. 69, no. 3, p. 036131, 2004.